Joshua Schnellenberger

Cherry Blossom Prediction Narrative

The Cherry Blossom Prediction Competition tasks participants with predicting the upcoming bloom dates for the cherry blossom trees at five different locations around the world. I chose to tackle the problem with a rather simple approach, using temperature and location as my only predictors. Most of the data I used was loaded with the help of the demonstration files provided to participants through GitHub. I built my model using a training dataset containing the historic bloom dates and temperatures for as many years as were available, going back to 2000 for each of the five locations. I chose only to go back to 2000 so as to not chase past trends in temperature. While my set of predictors was rather simple, ultimately using both the average and cumulative temperatures as well as location, I wanted to be able to use the insights from multiple different model types when making my predictions to account for the uncertainty when generating predictions based on somewhat limited data. I decided to implement a stacking approach that used linear regression, random forest, and XGBoost (Extreme Gradient Boost) models. Each model was trained individually using an LOOCV approach to prevent overfitting and hopefully improve predictive power. These models were then combined to create the final stacked or "meta" model, which is a linear combination of the three separate models. The XGBoost model turned out to be the most significant when predicting the bloom date (corresponding p-value: ~0). While it is more difficult to interpret the resulting coefficients of this stacked model than it would be for a simple linear regression model, I believe the resulting final model indicates that the combination of three essential methods can account for most of the variation in bloom date (Adjusted R-Squared: .8665).